Выявление факторов риска острых нарушений мозгового кровообращения на основе интеллектуального анализа историй болезни*

В.В. Донитова, Д.А. Киреев, Б.А. Кобринский, И.В. Смирнов, Е.В. Титова

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. Инсульт занимает второе место в мире среди причин смертности и третье место среди причин инвалидности и смертности вместе взятых. При этом среди факторов риска возникновения инсульта имеются потенциально управляемые, т.е. возможна профилактика данного заболевания. Выявление ранее неизвестных модифицируемых факторов риска инсульта или проверка значимости известных факторов являются актуальными задачами, которые можно решать на основе ретроспективного анализа историй болезни пациентов с этим заболеванием. В работе представлен подход к выявлению факторов риска острых нарушений мозгового кровообращения из текстов историй болезни с применением методов обработки естественного языка и машинного обучения. Предложенный подход позволил определить факторы риска инфаркта мозга и транзиторной ишемической атаки у пациентов одной из федеральных клиник. Выявленные факторы в целом согласуются с полученными в других исследованиях.

Ключевые слова: *факторы риска, инсульт, извлечение информации из текстов, машинное обучение, истории болезни.*

DOI: 10.14357/20790279230211

Введение

В большинстве развитых стран, где длительное время уделяли большое внимание совершенствованию мер предотвращения заболеваний, а не только их лечению, произошли значительные положительные сдвиги в показателях смертности от сосудистых заболеваний, что в результате и определило большой отрыв в продолжительности жизни и эффективности борьбы с основными хроническими неинфекционными заболеваниями (ХНИЗ) [1]. Возможности динамического контроля за факторами риска развития ХНИЗ с использованием интеллектуальной системы рассмотрены на примере кардио- и церебральной сосудистой патологии в [2]. Но важным и недорешенным вопросом является выявление таких факторов в электронных медицинских картах (ЭМК), в частности в историях болезни стационарных больных, что является сложной задачей ввиду слабой структурированности медицинских документов, синонимии, произвольного использования сокращений и др.

Особое значение своевременное выявление факторов риска имеет для социально значимых заболеваний, одним из которых является инсульт. По последним данным инсульт занимает второе место в мире среди причин смертности и третье место среди причин инвалидности и смертности вместе взятых [3]. При проведении системного анализа ОНМК, по данным за 1990-2016 гг., обнаружено, что в 87,9% случаев ишемического инсульта и 89,5% геморрагического инсульта имели место потенциально модифицируемые факторы риска [4]. Показано, что контроль модифицируемых факторов позволяет снизить риск инсульта [5,6]. Поэтому выявление модифицируемых (управляемых) факторов риска является актуальной задачей.

Такой анализ возможен с использованием технологий и методов искусственного интеллекта для извлечения знаний из текстов на естественном языке в массиве электронных медицинских карт. Исследования в этом направлении важны, так как нет чётких рекомендаций, какой из алго-

* Работа выполнена при поддержке РФФИ в рамках научного проекта № 19-29-01090 мк.

ритмов будет работать лучше в той или иной ситуации в такой слабо структурированной области как медицина [7].

В настоящей работе решается задача выявления риск-факторов острых нарушений мозгового кровообращения при помощи извлечения потенциальных факторов из текстовых (неструктурированных) историй болезни с последующим выделением значимых показателей и их комбинаций методами машинного обучения.

1. Подходы к интеллектуальному анализу разнородных медицинских данных

Данные о состоянии здоровья людей хранятся в ЭМК в разных форматах. Это, как правило, структурированные и неструктурированные текстовые данные, а также изображения. В зависимости от степени структурированности данных применяются различные подходы к их обработке. В работе [8] представлен один из последних обзоров работ по применению искусственного интеллекта (ИИ) к обработке данных ЭМК. В частности, авторы отмечают недавние работы о применении ИИ для задачи прогнозирования появления таких заболеваний как сахарный диабет и гипертония, а также для предсказания риска остановки сердца. Особенно отмечена необходимость разработки специальных процедур и политик для корректного сбора и обработки медицинских данных с помощью ИИ.

В российском исследовании на данных 2131 человека методом машинного обучения с использованием нейронной сети осуществлялось прогнозирование исходов и рисков сердечно-сосудистых заболеваний у пациентов с артериальной гипертонией [9]. В работе Баранова А.А. и др. [10] в ходе комплексного анализа медицинских данных применялись методы извлечения информации из текстов и машинного обучения на основе деревьев решений. В результате исследования была показана целесообразность использования подобных комплексных решений для анализа медицинских текстов и данных для выделения наиболее значимых признаков для предсказания диагноза.

Обработка текстовой составляющей электронных медицинских карт является отдельной задачей. Проблемы в этой области хорошо описаны, например, в работе M. Tayefi и др. [11]. Основной и главной проблемой здесь является специфика медицинских текстов, из-за чего такие задачи как токенизация, раскрытие сокращений и исправление ошибок становятся сложнее, потому что они требуют либо значительной модификации существующих решений, либо создания совершенно новых решений, уникальных для каждого исследования.

Большинство современных исследований в области интеллектуального анализа медицинских текстов основаны на применении машинного обучения по размеченным корпусам текстов, в связи с чем роль последних сильно возросла. Стали появляться такие корпуса и для русскоязычных медицинских текстов. В работе [12] представлен один из первых русскоязычных размеченных корпусов клинических текстов, а также методы извлечения и связывания различных медицинских сущностей из текстов. В одной из последних работ P. Blinov и др. [13] для оценки работы методов обработки текстов медицинской тематики представлен русскоязычный набор тестов RuMedBench. В данный набор входят тесты по их классификации, ответу типа да/нет на контекстные вопросы, созданию выводов (Natural Language Inference, NLI) и извлечение именованных сущностей (NER) для текстов медицинской тематики на русском языке. Параллельно происходит и развитие т. н. нейросетевых языковых моделей. Так, в работе A. Yalunin и др. [14] была представлена модель RuBioRoBERTa, которая представляет собой модель RoBERTa [15], дообученную на текстах биомедицинской тематики на русском языке. Такая модель показала лучшее качество на наборе тестов RuMedBench [13].

При анализе медицинских структурированных данных важна задача выявления причинно-следственных отношений в данных. Для решения такого рода используются логические (индуктивные) методы машинного обучения. Например, в работе [16] описан один из таких методов – AQJSM, который применялся, например, для исследования связи агрессивности с различными личностными особенностями [17]. Он применялся также и в настоящей работе для выявления значимых факторов риска острых нарушений мозгового кровообращения.

2. Схема исследования

Данное исследование проводилось по следующей схеме:

1. Создание обезличенной выборки ЭМК (историй болезни) для 3-х групп пациентов: с ишемическим инсультом (инфаркт мозга), транзиторной ишемической атакой (ТИА) и контрольной группы (с отсутствием в анамнезе любой из форм инсульта).
2. Формирование выборки историй болезни для разметки, обучения и тестов.
3. Формирование перечня показателей, которые рассматривались в качестве потенциальных

факторов риска острых нарушений мозгового кровообращения.
4. Разметка ранее отобранной выборки текстов историй болезни экспертами.
5. Обучение методов извлечения потенциальных факторов риска на размеченной выборке.
6. Извлечение потенциальных факторов риска из всех имеющихся текстов обученными ранее методами.
7. Группирование потенциальных факторов по историям болезни пациентов с различными диагнозами. Нормализация классов по количеству экземпляров путем удаления лишних экземпляров.
8. Применение методов машинного обучения к полученной ранее выборке потенциальных факторов для извлечения значимых факторов риска острых нарушений мозгового кровообращения.
9. Анализ и интерпретация полученных результатов экспертами-медиками.

3. Исходные данные для выявления потенциальных факторов риска

3.1. Материал исследования

В качестве материала для исследования факторов риска выступили обезличенные ЭМК пациентов Федерального научно-клинического центра Федерального медико-биологического агентства (ФНКЦ ФМБА). Материал исследования первоначально включал ЭМК пациентов с ишемическим инсультом (инфаркт мозга), геморрагическим инсультом, транзиторной ишемической атакой (ТИА) и контрольную группу, в которой отсутствовали диагнозы любой из форм инсульта. Однако вследствие крайне малого числа ЭМК с диагнозом геморрагического инсульта, последующий анализ проводился только в отношении инфаркта мозга и ТИА.

Исследуемая выборка состояла из 1235 ЭМК, принадлежащих 328 пациентам, среди которых 220 (67%) – с диагнозом ишемический инсульт, 108 (33%) – транзиторная ишемическая атака. Эти пациенты подходили по нескольким критериям отбора. У каждого пациента было не менее двух госпитализаций, при последней из которых поставлен целевой диагноз (ишемический инсульт или ТИА). В качестве еще одного условия был определен временной интервал с 01.01.2009 года по 31.05.2019 года, в течение которого пациентам оказывалась медицинская помощь в ФНКЦ ФМБА. Такие границы обусловлены тем, что в течение последних 10 лет методы диагностики нарушений мозгового кровообращения и подходы к ведению историй болезни в ФНКЦ ФМБА практически не изменялись.

В контрольную выборку вошли пациенты, которые обращались в ФНКЦ ФМБА по медицинским показаниям или для плановых обследований, операций. Ни у одного из участников контрольной группы в анамнезе или в момент госпитализации не было ишемического инсульта или ТИА. Контрольная группа состояла из 1654 ЭМК, принадлежащих 500 пациентам.

Все пациенты, в соответствии с временем постановки им целевого диагноза, были разделены на 3 группы: пациенты моложе 40 лет, от 40 до 60 и старше 60 лет. Среди больных ишемическим инсультом большую часть составляли пациенты старше 60 лет – 177 (80,5%), среди больных ТИА большую часть составляли пациенты в возрасте от 40 до 60 лет – 53 (49,1%).

3.2. Показатели, исследуемые в качестве потенциальных факторов риска инсульта

На основе анализа литературных данных был сформирован список диагнозов и признаков, подлежащих извлечению из электронных медицинских карт пациентов как потенциальные факторы риска инсульта. Он включал 6 сущностей: артериальная гипертония (АГ), ишемическая болезнь сердца (ИБС), головокружение, сахарный диабет (СД), аритмия, стеноз магистральных артерий головы и шеи (стеноз МАГ). Кроме того, исследовалась информация о лабораторном показателе «Скорость оседания эритроцитов» (СОЭ), повышение которого может свидетельствовать о наличии генерализованного воспаления в организме, что может указывать на его роль как триггера при инсульте [18]. Измерение СОЭ обычно происходит с помощью одного из методов: по Вестергрену или по Панченкову. Так как нормальные значения для этих методик различаются, это было необходимо учитывать при выделении показателей, указывающих на повышение СОЭ. По Вестергрену для женщин моложе 50 лет, повышением СОЭ считаются значения больше 20 мм/ч, для женщин старше 50 - больше 30 мм/ч, для мужчин моложе 50 и старше 50 - больше 15 мм/ч и больше 20 мм/ч. По Панченкову значения СОЭ отмечаются как повышенные при значениях больше 12 мм/ч для женщин и больше 10 мм/ч для мужчин.

3.3 Разметка текстов историй болезни

Из сформированной на первом шаге выборки ЭМК была отобрана 661 история болезни, необходимая для разметки с целью обучения и проверки методов извлечения информации из текстов.

В текстах историй болезни размечались перечисленные выше потенциальные факторы риска в

Табл. 1

Количество размеченных факторов в текстах ИБ

Название потенциального фактора	Количество вхождений
Артериальная гипертония	1064
Стеноз	788
Аритмия	389
Ишемическая болезнь сердца	814
Головокружение	213
Сахарный диабет	342
Скорость оседания эритроцитов	289

соответствии со схемой, описанной в наших предыдущих публикациях [19,20]. В табл. 1 представлено количество размеченных вхождений каждого фактора.

4. Извлечение целевой информации из текстов историй болезни

Для извлечения потенциальных факторов риска из текстов ЭМК использовались методы, основанные на машинном обучении и описанные ранее в работе [19], с некоторой модификацией:

– для метода, основанного на глубоком обучении, рассматривалась дополнительная модель RuBioRoBERTa [14], предобученная на текстах медицинской тематики. Следствием использования этой модели явилось повышение точности извлечения признаков;
– для оценки результатов работы методов использовалась библиотека seqeval [21], которая специализируется на задачах маркировки последовательностей. Использование этой библиотеки позволило более точно оценить качество работы методов и дало возможность сравнения методов друг с другом.

В табл. 2 приведены результаты оценки качества работы модифицированных методов на размеченной тестовой выборке текстов. Для работы методов, основанных на CRF и BERT, были подобраны гиперпараметры для максимизации оценки F_1 на 100 итерациях. Для каждого фактора был выделен метод, лучше всего извлекающий этот фактор по оценке F_1.

Как видно из табл. 2, для 5 из 7 исследуемых факторов лучше всего показал себя метод, основанный на RuBERT и RuBioRoBERTa.

Извлечение потенциальных факторов риска инсульта выполнялось на всей выборке исходных данных, описанной в Разделе 3.1. Для извлечения каждого потенциального фактора на новых неразмеченных данных использовался метод, показавший лучшее качество на тестовой выборке текстов для данного фактора.

Табл. 2

Результаты оценки качества методов извлечения информации из текстов, значения в %

Основа метода / Фактор		АГ	Стеноз МАГ	Аритмия	ИБС	Головокружение	СД	СОЭ
Правила	Precision	98,80	89,62	55,81	52,94	91,41	91,12	60,65
	Recall	85,42	29,57	50,64	71,16	84,98	92,98	65,28
	F1	91,62	44,47	53,10	60,71	88,08	92,04	62,88
	Accuracy	98,77	95,79	93,37	95,97	99,12	99,28	99,28
CRF	Precision	95,92	**94,07**	73,81	77,06	93,94	96,19	**69,79**
	Recall	92,68	**91,19**	74,40	74,65	88,57	87,83	**96,79**
	F1	94,27	**92,61**	74,10	75,84	91,18	91,82	**96,79**
	Accuracy	99,41	**99,18**	95,78	97,81	99,52	99,21	**99,44**
BERT (RuBERT)	Precision	**97,00**	91,40	72,73	77,96	**93,75**	93,26	72,20
	Recall	**97,47**	91,88	75,82	83,06	**94,24**	92,22	70,00
	F1	**97,30**	91,64	74,24	80,43	**94,00**	92,73	71,08
	Accuracy	**99,76**	99,15	96,59	97,43	**99,48**	99,37	99,06
BERT (RuBioRoBERTa)	Precision	96,31	92,08	**71,71**	**82,9**	91,53	**92,86**	73,42
	Recall	95,74	90,48	**82,30**	**83,38**	91,05	**94,27**	75,65
	F1	96,03	91,27	**76,64**	**83,14**	91,29	**93,56**	74,52
	Accuracy	99,60	99,22	**96,38**	**97,85**	99,21	**99,27**	99,09

Табл. 3

Процентные отношения встречаемости потенциальных факторов риска для каждого из диагнозов

Потенциальный фактор риска или комбинация факторов		**Процентное отношение встречаемости записей о потенциальном факторе риска к общему количеству записей с диагнозом «Инфаркт мозга»**	**Процентное отношение встречаемости записей о потенциальном факторе риска к общему количеству записей с диагнозом «ТИА»**
Аритмия		56,82%	67,59%
СД		19,55%	10,19%
ИБС		68,64%	87,04%
АГ		92,73%	91,67%
Головокружение		52,27%	69,44%
Стеноз МАГ		71,82%	61,11%
АГ и стеноз МАГ		55,00%	48,15%
АГ и аритмия		44,55%	53,70%
АГ и СД		14,09%	8,33%
АГ и ИБС		58,18%	78,70%
Стеноз МАГ и аритмия		29,55%	35,19%
Стеноз МАГ и диабет		8,18%	2,78%
Стеноз МАГ и ИБС		34,55%	49,07%
Аритмия и СД		6,82%	6,48%
Аритмия и ИБС		32,27%	50,93%
СД и ИБС		6,82%	5,56%
АГ, стеноз МАГ и аритмия		22,73%	28,70%
АГ, стеноз МАГ и СД		4,55%	2,78%
АГ, стеноз МАГ и ИБС		28,18%	40,74%
АГ, аритмия и СД		4,55%	4,63%
АГ, аритмия и ИБС		26,82%	45,37%
АГ, СД и ИБС		5,91%	5,56%
Стеноз МАГ, аритмия и СД		2,27%	1,85%
Стеноз МАГ, аритмия и ИБС		15,00%	30,56%
Стеноз МАГ, СД и ИБС		3,18%	2,78%
Аритмия, СД и ИБС		1,82%	4,63%
Возраст	<40	0,45%	8,33%
	40-60	19,09%	49,07%
	>60	80,45%	42,59%

Предварительно проведенный частотный анализ показал, что при инфаркте мозга, по сравнению с ТИА, значительно чаще встречаются такие факторы как артериальная гипертония (32,59% против 15,81%), стеноз магистральных сосудов шеи и головы (30,56% против 12,77%), аритмии (28,54% против 16,67%), ишемическая болезнь сердца (28,49% против 17,74%) и сахарный диабет (35,25% против 9,02%). Это указывает на больший или меньший негативный вклад факторов риска в сравнительном анализе этих двух форм острого нарушения мозгового кровообращения, обусловленных спазмом сосудов мозга.

Показано, что при сочетании двух и более неблагоприятных факторов вероятность возникновения инсульта увеличивается [22]. Поэтому нами были построены комбинации из двух и трех факторов, включающих уже отобранные признаки (артериальная гипертония, стеноз магистральных сосудов, аритмия, сахарный диабет, ИБС). Особое внимание уделено высокой встречаемости автоматически выявленных комбинаций факторов риска для инфаркта мозга, по сравнению с ТИА, для стеноза МАГ и СД (40,00% против 6,67%) и для АГ, стеноза МАГ и СД (41,67% против 12,50%).

Более существенным является анализ отношения количества записей, в которых встречается тот или иной фактор или комбинация факторов риска, к общему количеству записей с тем или иным диагнозом (табл. 3).

Оценка частоты встречаемости каждого из потенциальных факторов риска в процентном отношении при каждом из диагнозов (табл. 3) позволяет сделать предварительное заключение о возможном значении рассматриваемых факторов

и их сочетаний. Преобладающий возраст старше 60 лет практически в 2 раза чаще встречается при инфаркте мозга, чем при ТИА, что подтверждает известный факт. ИБС и аритмия несколько чаще приводят к ТИА, при которой и более высокая частота головокружения, что имеет соответствующее содержательное объяснение. Стеноз магистральных сосудов головы и шеи и сахарный диабет чаще встречаются при инфаркте мозга, в частности и потому, что контингент этих больных в целом старше. При рассмотрении комбинаций факторов риска можно видеть, что при инфаркте мозга значительно чаще, чем при ТИА, имеют место такие сочетания как АГ + СД, стеноз МАГ + СД, АГ + стеноз МАГ + СД. В то время как при ТИА чаще встречаются следующие сочетания: АГ + ИБС, стеноз МАГ + аритмия, аритмия + ИБС, АГ + стеноз МАГ + ИБС, АГ + аритмия + ИБС, стеноз МАГ + аритмия + ИБС. Большинство этих комбинаций факторов риска отвечает содержательным представлениям. Частично оно объясняется коморбидностью патологии, нарастающей с возрастом (включая как увеличение тяжести ранее возникших патологических проявлений и заболеваний, так и появление новых хронических болезней).

5. Выявление факторов риска нарушения мозгового кровообращения

Следующий этап исследования был направлен на переход от анализа частотных характеристик извлеченных из историй болезни потенциальных факторов риска к выявлению значимых факторов риска нарушения мозгового кровообращения методами машинного обучения. Учитывая существующие явные и скрытые связи факторов (каузальные, ассоциативные), совместно влияющих на развитие патологических процессов, существовала необходимость рассмотрения каждого из потенциальных факторов риска в отдельности и в разных комбинациях.

Для выявления факторов риска болезней, связанных с нарушением мозгового кровообращения, были выбраны следующие методы: AQJSM [16], ориентированный на поиск причинно-следственных связей, регрессионная модель и дерево принятия решений. Данные методы были выбраны в связи с большими возможностями интерпретации, в отличии от, например, моделей глубокого обучения, интерпретация которых сильно затрудняется большим количеством параметров и слоев.

Рассмотрим далее результаты применения выбранных методов.

Метод AQJSM не выдал гипотез о факторах риска для инсульта, что можно объяснить большим сходством исследуемых целевых групп с контролем.

Модель логистической регрессии использует все данные, подаваемые на вход, но для интерпретации были выделены 8 наиболее значимых признаков. В табл. 4 представлены результаты анализа для пациентов с диагнозом «Инфаркт мозга». В табл. 5 – результаты анализа для пациентов с диагнозом «ТИА». Если признак отмечен символом «+», то это означает, что он может рассматриваться как фактор риска для данного диагноза (заболевания), символ «-» указывает на отсутствие влияния (или слабое влияние). Как видно, в качестве факторов риска для ишемического инсульта выделены: артериальная гипертония, аритмия, стеноз магистральных артерий головы и шеи, сахарный диабет, ишемическая болезнь сердца и возраст старше 60 лет. В то время как для ТИА характерны следующие факторы (в отдельности и в комбинации): АГ, головокружение, АГ+ИБС, стеноз МАГ+ИБС, стеноз МАГ+аритмия+ИБС, при возрасте до 60 лет.

Для данного метода можно утверждать, что чем больше у пациента факторов с плюсом, тем больше вероятность, что обученная модель логи-

Табл. 4

Результаты применения регрессионной модели для пациентов с диагнозом «Инфаркт мозга»

Фактор		**Инфаркт мозга**
Возраст	<40 лет	-
	40–60 лет	-
	>60 лет	+
Стеноз МАГ и ИБС		-
Аритмия		+
АГ		+
Стеноз МАГ		+
СД		+
ИБС		+
АГ, стеноз МАГ и ИБС		-

Табл. 5

Результаты применения регрессионной модели для пациентов с диагнозом «ТИА»

Фактор		**Транзиторная ишемическая атака**
АГ и ИБС		+
Головокружение		+
Возраст	<40 лет	+
	40-60 лет	+
	>60 лет	-
Стеноз МАГ, аритмия и ИБС		+
Стеноз МАГ и аритмия		-
АГ		+
Стеноз МАГ		-
Стеноз МАГ и ИБС		+
Аритмия и ИБС		-

стической регрессии предскажет истинный риск для соответствующего диагноза.

Применение метода построения дерева решений продемонстрировало следующие результаты: для «Инфаркт мозга», в отличие от «ТИА», выявлено присутствие такого показателя как СОЭ, который рассматривается как возможный триггер развития инсульта. Невыявление СОЭ для «ТИА» указывает на отсутствие воспалительного процесса (маркером которого является СОЭ), что соответствует клинической картине ТИА.

6. Обсуждение результатов

Метод логистической регрессии показал наиболее адекватные результаты. С помощью метода логистической регрессии в качестве факторов риска для ишемического инсульта выделены: артериальная гипертония, аритмия, стеноз магистральных артерий головы и шеи, сахарный диабет, ишемическая болезнь сердца и возраст старше 60 лет. Их частота у пациентов с ишемическим инсультом в анализируемой выборке: артериальная гипертония 93% случаев, стеноз магистральных артерий головы – 72%, ишемическая болезнь сердца – 69%, аритмия – 57%, головокружение – 52%, сахарный диабет – 20%.

Для ТИА характерны следующие факторы (в отдельности и в комбинации): АГ, головокружение, АГ+ИБС, стеноз МАГ+ИБС, стеноз МАГ+ аритмия+ИБС, при возрасте до 60 лет.

Помимо факторов риска, которые могут привести к развитию острого нарушения мозгового кровообращения, с помощью метода дерева принятия решений, был выделен маркер, который может указывать на возникновение нарушения мозгового кровоснабжения вследствие ассоциации с инфекциями, – «повышение скорости оседания эритроцитов (СОЭ)» [23]. Повышение СОЭ было зафиксировано у 20,9% пациентов.

Обратимся к сравнительному анализу факторов риска, полученных в процессе извлечения знаний из текстов в настоящем исследовании, с некоторыми аналогичными отечественными и зарубежными исследованиями. В исследовании INTERSTROKE [24], охватывающем практически все континенты, гипертония была значимо связана со всеми формами инсульта во всех регионах. В немецком исследовании [25] авторы указывают на наличие классических факторов риска, таких как гипертония и сахарный диабет, что подтверждает и наше исследование. В Японии гипертония также постоянно ассоциируется с повышенным риском инсульта независимо от возрастной категории, а сахарный диабет связан с повышенным риском инсульта в возрасте 60–74 лет, но не у лиц в возрасте ≥ 75 лет [26]. В российском исследовании [27] аритмия продемонстрировала встречаемость более, чем в 50% случаев. Таким образом, артериальная гипертония и сахарный диабет, как ведущие факторы риска ишемического инсульта, были подтверждены при анализе массива историй болезни ФНКЦ ФМБА. Другие рассмотренные факторы характеризуются определенной вариабельностью при инфаркте мозга и ТИА.

Заключение

В работе продемонстрирована возможность извлечения из неструктурированных текстов историй болезни потенциальных факторов риска нарушения кровообращения мозга и выявления среди них наиболее значимых в плане реализации предрасположения к заболеванию. Для этого использовались современные методы извлечения информации из текстов, включая нейросетевые языковые модели типа трансформер. Экспериментальная проверка показала достаточно высокое качество работы указанных методов. Для выявления и оценки влияния факторов

риска на развитие ишемического инсульта и ТИА использовался метод логистической регрессии, метод построения дерева решений и AQJSM. Лучшие результаты показал метод логистической регрессии. С использованием этого метода для ишемического инсульта выявлены такие факторы как возраст старше 60 лет, артериальная гипертензия, аритмия, стеноз магистральных артерий головы и шеи, сахарный диабет и ишемическая болезнь сердца. Для ТИА наиболее значимыми факторами риска являются: возраст моложе 60 лет, наличие артериальной гипертензии и головокружения. Кроме того значительное влияние на возникновение ТИА оказывают следующие комбинации: артериальная гипертензия и ишемическая болезнь сердца, стеноз МАГ и ишемическая болезнь сердца, а также наличие стеноза МАГ, аритмии и ишемической болезни сердца одновременно.

Выявленные на российской выборке с использованием методов машинного обучения факторы риска и их комбинации корреспондируют, по наиболее известным факторам риска, с результатами других исследований, выполненных в странах Европы, Азии и Африки. В то же время, полученные результаты продемонстрировали некоторые отличия, для подтверждения которых необходимы дальнейшие исследования на расширенной российской выборке.

Литература

1. *Гусев А.В., Кузнецова Т.Ю., Корсаков И.Н.* Искусственный интеллект в оценке рисков развития сердечно-сосудистых заболеваний //Журнал телемедицины и электронного здравоохранения. 2018. №. 3. С. 85-90.
2. *Кобринский Б.А., Кадыков А.С., Полтавская М.Г., Благосклонов Н.А., Ковелькова М.Н.* Принципы функционирования интеллектуальной системы динамического контроля факторов риска и формирования рекомендаций по здоровьесбережению // Профилактическая медицина. 2019. Т.22. №5. С.78- 84.
3. *Feigin V.L. et al.* Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019 // The Lancet Neurology. 2021. Vol. 20. №. 10. P. 795-820.
4. *Johnson C.O. et al.* Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016 //The Lancet Neurology. 2019. Vol. 18. №. 5. P. 439-458.
5. *Boehme A.K., Esenwa Ch., Elkind M.S.V.* Stroke Risk Factors, Genetics, and Prevention. Circulation Research. 2017. Vol. 120. P. 472–495.
6. *Chen J. et al.* Stroke Risk Factors of Stroke Patients in China: A Nationwide Community-Based Cross-Sectional Study //International Journal of Environmental Research and Public Health. 2022. Vol. 19. №. 8. P. 4807.
7. *Швец Д.А., Поветкин С.В.* Сравнительный обзор использования методов машинного обучения для прогнозирования сердечно-сосудистого риска // Вестник новых медицинских технологий. Электронное периодическое издание. 2020. № 5. С.74-82.
8. *Lee S., Kim H.S.* Prospect of artificial intelligence based on electronic medical record // Journal of Lipid and Atherosclerosis. 2021. Vol. 10. №. 3. P. 282.
9. *Невзорова В.А., Плехова Н.Г., Присеко Л.Г., Черненко И.Н., Богданов Д.Ю., Мокшина М.В., Кулакова Н.В.* Методы машинного обучения в прогнозировании исходов и рисков сердечно-сосудистых заболеваний у пациентов с артериальной гипертензией (по материалам ЭССЕ-РФ в Приморском крае) // Российский кардиологический журнал. 2020. № 3. С. 10-16.
10. *Баранов А.А. и др.* Технологии комплексного интеллектуального анализа клинических данных // Вестник Российской академии медицинских наук. 2016. Т. 71. №. 2. С. 160-171.
11. *Tayefi M. et al.* Challenges and opportunities beyond structured data in analysis of electronic health records //Wiley Interdisciplinary Reviews: Computational Statistics. 2021. Vol. 13. №. 6. P. e1549.
12. *Shelmanov A.O., Smirnov I.V., Vishneva E.A.* Information extraction from clinical texts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. 2015. Vol. 1. P. 537-549.
13. *Blinov P. et al.* RuMedBench: A Russian Medical Language Understanding Benchmark // Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science. 2022. Vol. 13263. P. 383- 392.
14. *Yalunin A., Nesterov A., Umerenkov D.* RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. 2022. URL: https://arxiv.org/abs/2204.03951.
15. *Liu Y. et al.* Roberta: A robustly optimized bert pretraining approach. 2019. URL: https://arxiv.org/abs/1907.11692.
16. *Панов А.И.* Выявление причинно-следственных связей в данных психологического тестирования логическими методами // Искусственный интеллект и принятие решений. 2013. №. 1. С. 24-32.

17. *Чудова Н.В., Панов А.И.* Извлечение причинно-следственных отношений из данных психологического исследования на материале изучения агрессивности //Искусственный интеллект и принятие решений. 2016. №. 4. С. 38- 46.
18. *Самойлова Е.М., Юсубалиева Г.М., Белопасов В.В., Екушева Е.В., Баклаушев В.П.* Инфекции и воспаление в развитии инсульта // Журнал неврологии и психиатрии им. С.С. Корсакова. Спецвыпуски. 2021. Т. 121. № 8. С.11- 21.
19. *Донитова В. В. и др.* Методы обработки естественного языка для извлечения факторов риска инсульта из медицинских текстов // Труды ИСА РАН. 2021. Т. 71. №. 4. С. 93.
20. *Благосклонов Н.А., Донитова В.В., Киреев Д.А., Кобринский Б.А., Смирнов И.В.* Лингвистический анализ историй болезни для выявления факторов риска инсульта // Труды ИСА РАН. 2020. Т. 70. № 3. С. 76- 86.
21. *Nakayama H.* Seqeval: A python framework for sequence labeling evaluation, Available at: https://github. com/chakki-works/seqeval (дата обращения 15.12.2022)
22. *Du X., McNamee R., Cruickshank K.* Stroke risk from multiple risk factors combined with hypertension: a primary care based case-control study in a defined population of northwest England // Annals of Epidemiology. 2000. Vol. 10. №. 6. P. 380-388.
23. *Sebastian S., Stein L.K., Dhamoon M.S.* Infection as a Stroke Trigger. Associations Between Different Organ System Infection Admissions and Stroke Subtypes // Stroke. 2019. Vol. 50. P. 2216 - 2218.
24. *O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al.* INTERSTROKE Investigators. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study // The Lancet. 2016. Vol. 388. №10046. P. 761-775.
25. *Thiele I., Linseisen J., Heier M., Holle R., Kirchberger I., Peters A. et al.* Time trends in stroke incidence and in prevalence of risk factors in Southern Germany, 1989 to 2008/09 // Scientific Reports. 2018. Vol. 8. №1. P. 1-8.
26. *Murakami K., Asayama K., Satoh M., Inoue R., Tsubota-Utsugi M. et al.* Miki Hosaka Risk Factors for Stroke among Young-Old and Old-Old Community-Dwelling Adults in Japan: The Ohasama Study // Journal of atherosclerosis and thrombosis. 2017. Vol. 24. P. 290-300.
27. *Усанова Т.А. и др.* Факторы риска ишемического инсульта // Современные проблемы науки и образования. 2020. №. 2. С. 133.

Донитова Виктория Владимировна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Научный сотрудник. Количество печатных работ: 20. Область научных интересов: извлечение знаний, интеллектуальные системы, системы поддержки принятия решений, экспертные системы.
E-mail: vdonitova@gmail.com (Ответственный за переписку).

Киреев Данил Алексеевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Программист. Количество печатных работ: 5. Область научных интересов: машинное обучение, глубокое обучение, активное обучение, обработка естественного языка, извлечение именованных сущностей. E-mail: kireev@isa.ru

Кобринский Борис Аркадьевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» Российской Академии Наук», г. Москва, Россия. Заведующий отделом. Доктор медицинских наук, профессор, заслуженный деятель науки РФ. Количество печатных работ: более 500 (в т.ч. 15 монографий и учебников). Область научных интересов: инженерия знаний, нечеткая логика, экспертные системы, интеллектуальные системы, системы поддержки принятия решений. E-mail: kba_05@mail.ru

Смирнов Иван Валентинович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» Российской Академии Наук», г. Москва, Россия. Заведующий отделом. Кандидат физико-математических наук, доцент. Количество печатных работ: 123. Область научных интересов: интеллектуальный анализ текстов и данных, обработка естественного языка. E-mail: ivs@isa.ru

Титова Елизавета Викторовна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Инженер-исследователь. Количество печатных работ: 5. Область научных интересов: обработка естественного языка, проблемно-ориентированные системы, экспертные системы, медицинские информационные системы. E-mail: elz.titova@gmail.com

Retrieving stroke risk factors based on intellectual analysis of electronic health records

V.V. Donitova, D.A. Kireev, B.A. Kobrinskii, I.V. Smirnov, E.V. Titova
Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. Stroke is the world's second leading cause of death and the third leading cause of disability and death combined. Risk factors for stroke are potentially manageable so prevention of this disease is possible. Identification of previously unknown modifiable risk factors for stroke or testing the significance of known factors is an urgent task that should be solved based on a retrospective analysis of the electronic health records of patients with this disease. The paper presents an approach to identifying risk factors for acute cerebrovascular accidents from texts of case histories using natural language processing and machine learning methods. The proposed approach made it possible to identify risk factors for stroke and transient ischemic attack in patients of one of the Moscow clinics. The identified factors are consistent with those found in other studies.

Keywords: *risk factors, stroke, information extraction from texts, machine learning, electronic health records*
DOI: 10.14357/20790279230211

References

1. *Gusev A.V., Kuznetsova T.Yu., Korsakov I.N.* 2018. Iskusstvennyi intellekt v otsenke riskov razvitiya serdechno-sosudistykh zabolevanii [Artificial intelligence in assessing the risks of developing cardiovascular diseases]. Zhurnal telemeditsiny i elektronnogo zdravookhraneniya [Journal of Telemedicine and eHealth]. 3: 85-90.
2. *Kobrinskii B.A., Kadykov A.S., Poltavskaya M.G., Blagosklonov N.A., Kovelkova M.N.* 2019. Printsipy funktsionirovaniya intellektualnoi sistemy dinamicheskogo kontrolya faktorov riska i formirovaniya rekomendatsii po zdorovesberezheniyu [Principles of functioning of an intelligent system for dynamic control of risk factors and the formation of recommendations for health saving]. Profilakticheskaya meditsina [Preventive medicine]. 22 (5): 78- 84. doi: 10.17116/profmed20192205178.
3. *Feigin V.L. et al.* 2021. Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology. 20 (10): 795-820.
4. *Johnson C.O. et al.* 2019. Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology. 18 (5): 439-458.
5. *Boehme A.K., Esenwa Ch., Elkind M.S.V.* 2017. Stroke Risk Factors, Genetics, and Prevention. Circulation Research. 120: 472–495. doi: 10.1161/CIRCRESAHA.116.308398.
6. *Chen J. et al.* 2022. Stroke Risk Factors of Stroke Patients in China: A Nationwide Community-Based Cross-Sectional Study. International Journal of Environmental Research and Public Health. 19 (8): 4807.
7. *Shvets D.A., Povetkin S.V.* 2020. Sravnitelnyi obzor ispolzovaniya metodov mashinnogo obucheniya dlya prognozirovaniya serdechno-sosudistogo riska [Comparative review of the use of machine learning methods for predicting cardiovascular risk]. Vestnik novykh meditsinskikh tekhnologii. Elektronnoe periodicheskoe izdanie [Bulletin of new medical technologies. Electronic periodical.]. 5: 74-82. doi: 10.24411/2075-4094-2020-16711.
8. *Lee S., Kim H.S.* 2021. Prospect of artificial intelligence based on electronic medical record. Journal of Lipid and Atherosclerosis. 10 (3): 282.
9. *Nevzorova V.A., Plekhova N.G., Priseko L.G., Chernenko I.N., Bogdanov D.Yu., Mokshina M.V., Kulakova N.V.* 2020. Metody mashinnogo obucheniya v prognozirovanii iskhodov i riskov serdechno-sosudistykh zabolevanii u patsientov s arterialnoi gipertenziei (po materialam ESSE-RF v Primorskom krae) [Machine learning methods in predicting the outcomes and risks of cardiovascular diseases in patients with arterial hypertension (based on ESSE-RF materials in Primorsky Krai)]. Rossiiskii kardiologicheskii zhurnal [Russian Journal of Cardiology]. 3: 10-16. doi: 10.15829/1560-4071-2020-3-3751.

10. *Baranov A.A. et al.* 2016. Tekhnologii kompleksnogo intellektualnogo analiza klinicheskikh dannykh [Technologies of complex intellectual analysis of clinical data]. Vestnik Rossiiskoi akademii meditsinskikh nauk [Bulletin of the Russian Academy of Medical Sciences]. 71 (2): 160-171.
11. *Tayefi M. et al.* 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics. 13 (6): e1549.
12. *Shelmanov A.O., Smirnov I.V., Vishneva E.A.* 2015. Information extraction from clinical texts in Russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". 1: 537-549.
13. *Blinov P. et al.* 2022. RuMedBench: A Russian Medical Language Understanding Benchmark. Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science. 13263: 383- 392.
14. *Yalunin A., Nesterov A., Umerenkov D.* RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. 2022. URL: https://arxiv.org/abs/2204.03951.
15. *Liu Y. et al.* Roberta: A robustly optimized bert pretraining approach. 2019. URL: https://arxiv.org/abs/1907.11692.
16. *Panov A.I.* 2013. Vyyavlenie prichinno-sledstvennykh svyazei v dannykh psikhologicheskogo testirovaniya logicheskimi metodami [Identification of causal relationships in psychological testing data by logical methods]. Iskusstvennyi intellekt i prinyatie reshenii [Artificial intelligence and decision making]. 1: 24-32.
17. *Chudova N.V., Panov A.I.* 2016. Izvlechenie prichinno-sledstvennykh otnoshenii iz dannykh psikhologicheskogo issledovaniya na materiale izucheniya agressivnosti [Extraction of causal relationships from the data of psychological research on the material of the study of aggressiveness]. Iskusstvennyi intellekt i prinyatie reshenii [Artificial intelligence and decision making]. 4: 38- 46.
18. *Samoilova E.M., Yusubalieva G.M., Belopasov V.V., Ekusheva E.V., Baklaushev V.P.* 2021. Infektsii i vospalenie v razvitii insulta [Infection and inflammation in the development of stroke]. Zhurnal nevrologii i psikhiatrii im. S.S. Korsakova. Spetsvypuski [Journal of Neurology and Psychiatry. S. S. Korsakov. Special editions]. 121 (8): 11- 21.
19. *Donitova V.V. et al.* 2021. Metody obrabotki estestvennogo yazyka dlya izvlecheniya faktorov riska insulta iz meditsinskikh tekstov [Natural language processing models for extraction of stroke risk factors from electronic health records]. Trudy Instituta sistemnogo analiza rossiyskoy akademii nauk [Proceedings of the Institute for Systems Analysis of the Russian Academy of Science]. 71 (4): 93.
20. *Blagosklonov N.A. et al.* 2020. Lingvisticheskii analiz istorii bolezni dlya vyyavleniya faktorov riska insulta [Linguistic analysis of disease history for identifying stroke risk factors]. Trudy Instituta sistemnogo analiza rossiyskoy akademii nauk [Proceedings of the Institute for Systems Analysis of the Russian Academy of Science].70 (3): 75- 85.
21. *Nakayama H.* Seqeval: A python framework for sequence labeling evaluation, Available at: https://github. com/chakki-works/seqeval (дата обращения 15.12.2022)
22. *Du X., McNamee R., Cruickshank K.* 2000 Stroke risk from multiple risk factors combined with hypertension: a primary care based case-control study in a defined population of northwest England. Annals of Epidemiology. 10 (6): 380- 388
23. *Sebastian S., Stein L.K., Dhamoon M.S.* 2019 Infection as a Stroke Trigger. Associations Between Different Organ System Infection Admissions and Stroke Subtypes. Stroke. 50: 2216 - 2218. doi: 10.1161/STROKEAHA.119.025872
24. *O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al.* 2016. INTERSTROKE Investigators. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The Lancet. 388 (10046): 761 - 775. doi: 10.1016/S0140-6736(16)30506-2
25. *Thiele I., Linseisen J., Heier M., Holle R., Kirchberger I., Peters A. et al.* 2018 Time trends in stroke incidence and in prevalence of risk factors in Southern Germany, 1989 to 2008/09. Scientific Reports. 8 (1): 1- 8. doi: 10.1038/s41598-018-30350-8
26. *Murakami K., Asayama K., Satoh M., Inoue R., Tsubota-Utsugi M. et al.* 2017 Miki Hosaka Risk Factors for Stroke among Young-Old and Old-Old Community-Dwelling Adults in Japan: The Ohasama Study. Journal of atherosclerosis and thrombosis. 24: 290-300. doi: 10.5551/jat.35766
27. *Usanova T.A. et al.* 2020. Faktory riska ishemicheskogo insulta [Risk factors for ischemic stroke]. Sovremennye problemy nauki i obrazovaniya [Modern problems of science and education]. 2: 133.

Donitova V.V. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. E-mail: vdonitova@gmail.com

Kireev D.A. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. E-mail: kireev@isa.ru

Kobrinskii B.A. PhD, Professor. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. E-mail: kba_05@mail.ru

Smirnov I.V. PhD, Assoc. Professor. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. E-mail: ivs@isa.ru

Titova E.V. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, Russia. E-mail: elz.titova@gmail.com